Methods and/or system for selecting data sets

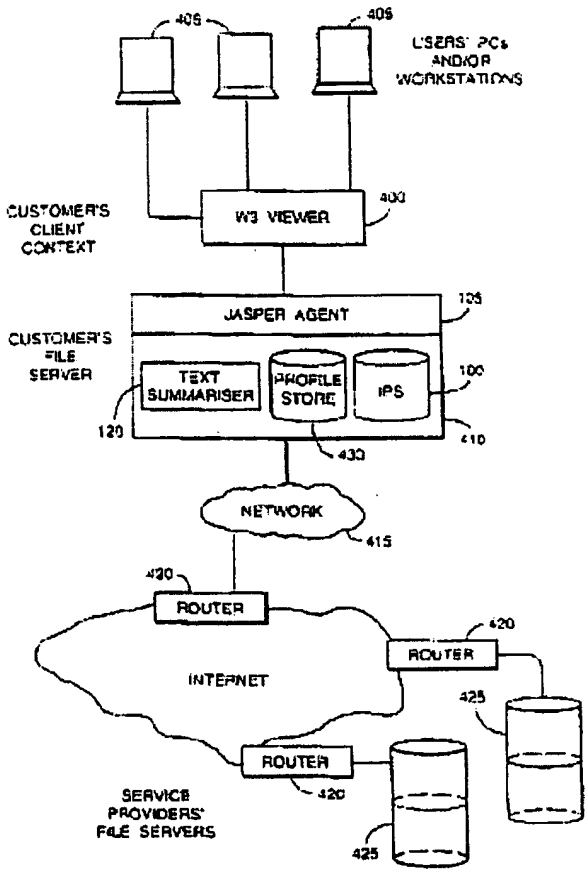
Patent number: CN1269897
Publication date: 2000-10-11
Inventor: DAVIES NICHOLAS JOHN (GB); WEEKS RICHARD (GB)
Applicant: BRITISH TELECOMM (GB)
Classification:
- International: G06F17/30
- european:
Application number: CN19980808771 19980828
Priority number(s): EP19970306878 19970904

Also published as:
WO9912108 (A1)
US6353827 (B1)
CA2302264 (A1)
AU742831 (B2)

Report a data error here

Abstract not available for CN1269897
Abstract of corresponding document: **US6353827**

Methods and apparatus for identifying associated key words in a data set. Associated key words are identified by a parser which firstly operates to extract key words from a data set. These key words are then analyzed by the parser to identify which key words, if any, have an association as determined by a predefined set of rules. These rules are grammatical and include, for example, two key words both being nouns that occur one after the other without intervening low value words. A similar rule applies to nouns followed by verbs but does not extend to verbs followed by nouns. These rules allow terms and phrases such as "information technology" and "wide area network" to be identified as associated key words rather than as individual and unrelated key words.



[19]中华人民共和国国家知识产权局

[51]Int.Cl[7]
G06F 17/30

[12]发明专利申请公开说明书

[21]申请号 98808771.5

[43]公开日 2000年10月11日

[11]公开号 CN 1269897A

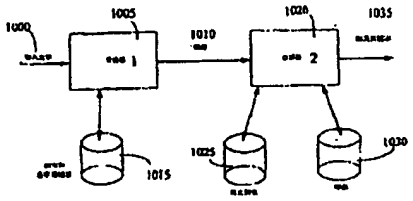
[22]申请日 1998.8.28 [21]申请号 98808771.5
[30]优先权
[32]1997.9.4 [33]EP [31]97306878.6
[86]国际申请 PCT/GB98/02611 1998.8.28
[87]国际公布 WO99/12108 英 1999.3.11
[85]进入国家阶段日期 2000.3.2
[71]申请人 英国电讯有限公司
地址 英国伦敦
[72]发明人 尼古拉斯·约翰·戴维斯
理查德·威克斯
[74]专利代理机构 永新专利商标代理有限公司
代理人 韩宏

权利要求书3页 说明书21页 附图页数7页

[54]发明名称 用于选择数据集的方法和/或系统

[57]摘要

用于在数据集(1000)中识别相关关键字(1035)的方法和设备。相关关键字由分析器(1020)识别,该分析器首先从数据集(1000)中提取关键字。这些关键字然后由分析器(1020)分析识别哪个是关键字,如果有的话,具有由预定规则集确定的相关性。这些规则是语法上的并包括例如都是名词一个随一个出现的两个关键字而没有插入的低值词。类似规则应用于由动词跟随名词,但不扩展到由名词跟随动词。这些规则允许术语或词组诸如"信息技术"和"广域网"作为相关关键字识别,而不是作为单独和无关的关键字。



ISSN1008-4274

00-03-02

权 利 要 求 书

1．用于确定在至少第一和第二数据集之间相似性量度的设备，所述设备包括：

i） 输入装置，用于接收至少所述第一和第二数据集；

ii） 处理装置，用于识别在至少第一数据集中的关键字集，该处理装置可以访问至少一个规则集，并通过使用所述至少一个规则集识别关键字集，该处理装置进一步确定所述相似性的量度；和

iii） 输出装置，输出相似的所述量度

其中所述规则集包括涉及在各个数据集中数据项目相对位置的规则，和其中所述处理装置通过比较由在第一数据集中的所述处理装置所识别的至少一个关键字集与包括或来自所述第二数据集的关键字集来确定相似性量度。

2．权利要求 1 中的设备进一步包括信息检索装置和数据存储器，所述第一数据集包括由所述信息检索装置从信息库中所检索出的数据，所述第二数据集包括在所述数据存储器中存储的关键字集。

3．权利要求 1 中的设备进一步包括信息检索装置，所述第一和第二数据集包括由所述信息检索装置从信息库中所检索出的数据，该处理装置识别所述第一和第二数据集的每个中的关键字集并通过比较各个关键字集确定相似性量度。

4．权利要求 2 中的设备，其中所述第二数据集定义目标数据集，该目标数据集用于由所述信息检索装置从所述信息库中检索数据，由此当所述相似性量度超过预定阈值时所述第一数据集被所述处理装置识别为包含所述目标数据集。

5．前面任何一个权利要求中的设备，其中在各个数据集中数据项目的所述相对位置包括至少两个在数据集中相互相关的潜在关键字相邻位置，该处理装置识别这种相邻潜在关键字，将其一同作为

在所识别的关键字集中的单一关键字。

6．权利要求 5 中的中的设备，其中所述至少一个规则集包括至少下列标准之一：

1） 由一个名词或一个预定标记集跟随一个名词；

2） 由一个名词或一个预定标记集跟随一个动词；

3） 由一个名词或一个预定标记集跟随一个形容词；和

4） 由一个名词或一个动词或另外一个预定标记集跟随一个预定标记集

只有当它们符合所述至少一个标准时该处理装置识别相邻潜在关键字，将其一同作为在所识别关键字集中的单一关键字。

7．权利要求 2 到 6 中的设备，其中所述数据存储器包括由所述处理装置从由所述信息检索装置从所述信息库所检索的多个数据集中识别的多个关键字集，其中所述处理装置根据对每对数据集所计算出的相似性量度来定义在所述数据集之间的多个关系。

8．一种确定第一和第二数据集之间相似性水平的方法，其中所述方法包括步骤：

i） 按照至少第一规则，在至少第一数据集中对所选择数据项目加上识别标签；

ii） 通过参照所述识别标签的出现或空缺识别潜在关键字集；

iii） 通过应用至少第二规选择两个或更多相邻潜在关键字集；

iv） 将每个所选择潜在关键字集分类作为单一关键字；

v） 产生关键字集，其中该关键字集包括每个被分类为单一关键字的潜在关键字集，和来自所识别的潜在关键字集的剩余关键字；和

vi） 将所产生的关键字集与包括或来自第二数据集的关键字集进行比较。

9. 权利要求 8 中的一种方法，其中所述第一规则涉及至少数据项目的部分语法分类。

10. 权利要求 8 或 9 中的一种方法，其中所述至少第二规则包括一个或更多来自下列组的规则：

1） 由一个名词或一个预定标记集跟随一个名词；

2） 由一个名词或一个预定标记集跟随一个动词；

3） 由一个名词或一个预定标记集跟随一个形容词；和

4） 一个名词或一个动词或另外一个预定标记集跟随一个预定标记集。

00-03-02

说　明　书

用于选择数据集的方法和/或系统

本发明涉及用于选择数据集的方法和/或系统，本发明在从例如信息库诸如使用因特网可访问的信息库中选择文献尤其有用。

因特网万维网是基于连接在一起的多个独立通信网络的公知通信系统。它提供了来自许多不同提供者的丰富信息源，但是这种丰富产生了访问特定信息的问题，因为没有集中监视和控制。

在 1982 年，科学、公司和技术信息的容量每 5 年翻番。到 1988 年，每 2.2 年翻番和到 1992 年每 1.6 年翻番。随着因特网和其它网络的扩展增加的速度将持续增长。这种网络生存的关键是管理信息和当他们需要时向用户提供他们想要的信息的能力。

可是，本发明不涉及提供用于搜索诸如万维网（W3）的系统的另一个工具：已经有许多工具了。他们往往随环球网覆盖面的不断增加和搜索引擎的改进而增加。

相反，本发明的实施例涉及下列问题：在 W3 上找到有用信息，如何能够存储得容易检索和如何能够使其它的可能用户对识别和通知的信息感性趣？

更具体地说，申请人的共同审查中申请 PCT/GB96/00132 提供了一种信息检索代理程序，称为 JASPER 代理程序，其用于从分布信息系统诸如 W3 中识别和检索信息。

它使用诸如分级凝聚群集（hierarchical agglomerative clustering）技术限定 W3 上存在的各种信息源之间的关系。可是，在这些限定的关系中可能出现不精确。这会导致具有不相类似的主题的文件聚集在一起。群集技术的本性在于一个不精确的群集文献然后能够增加为几个。

根据本发明的第一方面提供设备用于确定在至少第一和第二数据集之间相似性的量度，所述设备包括：

i） 输入装置，用于接收至少第一和第二数据集；

ii） 处理装置，用于识别在至少第一数据集中的一个关键字集，该处理装置可以访问至少一个规则集并通过使用所述至少一个规则集来识别关键字集，该处理装置进一步确定相似性的所述量度；和

iii） 输出装置，以输出相似性的所述量度

其中所述规则集包括关于各自数据集中数据项相对位置的规则，和其中所述处理装置通过比较由所述处理装置在第一数据集中识别的至少一个关键字集与包括或来自所述第二数据集的关键字集来确定相似性量度。

本发明的实施例允许在一个数据集中的两个或更多关键字相互相关，例如组成词组的关键字，因此数据集相似性比较中的精确性可以改善。

优选地，该设备进一步包括信息检索装置和数据存储器，所述第一数据集包括通过所述信息检索装置从数据库检索的数据，和所述第二数据集包括存储在所述数据存储器中的关键字集。例如，可能已经由用户提供，或存储在用户文档中的关键字集。

该规则集可能提供装置以识别数据集中的相邻项目，这些相邻项目可以一起视作单一关键字。这不仅需要位置信息，而且需要例如相邻项目上的语法检验，诸如下列一个或更多项：

1） 一个名词之后紧跟一个名词或预定标记集；

2） 一个动词之后紧跟一个名词或预定标记集；

3） 一个形容词之后紧跟一个名词或预定标记集；和

4） 一个预定标记集之后紧跟一个名词或一个动词或仍然一个预定标记集。

根据本发明的第二方面提供一种方法，确定第一和第二数据集之间相似性程度，其中所述方法包括步骤：

i） 按照至少第一规则，对在至少第一数据集中所选择的数据项目加识别标签；

ii） 通过参考所述识别标签的出现或空缺识别潜在关键字集；

iii） 通过应用至少第二规则选择两个或更多相邻的潜在关键字集；

iv） 将每个所选择的潜在关键字集分类作为单一关键字；

v） 产生关键字集，该关键字集包括作为单一关键字的每个分类的潜在关键字集和来自所识别的潜在关键字集的剩余关键字；

vi） 将所产生的关键字集与包括或来自第二数据集的关键字集进行比较。

例如，所述第一规则可以有利地至少部分涉及数据项目的语法类别。

所述至少第二规则可以包括一个或更多来自下列集的规则：

1） 一个名词之后紧跟一个名词或预定标记集；

2） 一个动词之后紧跟一个名词或预定标记集；

3） 一个形容词之后紧跟一个名词或预定标记集；和

4） 一个预定标记集之后紧跟一个名词或一个动词或仍然一个预定标记集。

与现有技术的系统和方法相比，在位于 W3 上和其它信息库上的文献、和其它形式信息中识别相关关键字改善了这些文献、和其它形式信息之间限定关系的精确性。

现在仅通过举例方式参照附图描述用于选择数据集的一种方法和/或系统，图中：

图 1 表示结合 Jasper 代理程序系统的一种信息访问系统；

图 2 以示意性格式表示由访问系统提供的存储过程；

图 3 表示用于图 1 的存储处理中智能页面存储的结构；

图 4 表示由访问系统提供的检索处理示意性格式；

图 5 表示图 2 的存储处理的流程图；

图 6、7 和 8 表示用于使用 Jasper 访问系统的三个信息检索处理的流程图；

图 9 表示使用群集技术产生的关键字网络，用于扩展和/或应用 Jasper 系统中的用户文档；

图 10 表示图 1 的 Jasper 代理程序的一部分，用于识别相关关键字。

本发明的实施例提供对诸如下面描述的 JASPER 代理程序的信息访问和信息检索系统的改进。随后提供的本发明实施例说明是针对该 JASPER 代理程序的说明。

可是，本发明不限于 JASPER 代理程序。它还在其它领域中应用，例如使用用户文档技术的信息系统和使用关键字检索和关键字搜索技术的信息系统。

一种信息访问系统

软件代理程序提供了应付分布式而非集中式计算机平台系统的公知方案。每个代理程序通常包括以自主方式代表实体（人或机器平台）与本地数据、或访问数据的装置一起执行一个任务或多个任务的功能，以支持该任务或多个任务。在本说明书中，在本发明实施例中用于存储或检索信息的代理程序为简单起见称为“Jasper 代理程序”，该词干来自“利用容易检索对存储页面联合访问 Joint Access to Stored Pages with Easy Retrieval”的首字母缩写。

假如在 W3 上有大量有用信息，最好避免将信息从源位置拷贝到本地服务器上。实际上，可以证明这种方案与环球网的整个意义相矛盾。因此，与其拷贝信息不如用 Jasper 代理程序只存储有关“元信息”。如下面所见，该元信息可以被认为是在信息本身更上层的、

关于这些信息而非这些实际信息。它可以包括例如关键字、概述、文献标题、通用资源定位器（URL）和数据及时间入口。当产生检索要求时，该元信息然后用于提供对实际信息的指针或“索引”。

大多数公知的 W3 用户（例如 Mosaic™和 Netscape™）提供一些装置用于存储关于用户感性趣页面的信息。典型地，这是通过允许用户产生有关特定 URL 名录的（可能是分层结构的）菜单。尽管菜单设施有用，当涉及相当大量的 W3 页面时它很快变得不方便使用。实质上，所提供的显示不足以允许收集可能需要的所有关于所存储信息的内容：用户只能提供命名该页面的字符串。象实际上有用的元信息诸如页面入口数据丢失一样，单一词组不足以（该名称）在各个环境对页面精确索引。

考虑关于药理学数据信息检索中使用基于知识系统（KBS）作为简单实例信息：在可以是任何 KBS 的不同环境下，检索信息或感性趣的药理学。除非仔细选择的名称提及所有三个方面，在一个或更多有用环境中将丢失信息。此问题与在 Unix（或其它的）文件系统中寻找包含所需要信息的文件相类似，如同 Jones, W. P. 在文章“人工存储模型的实际使用：存储扩展器个人文件编排系统”中所描述的，发表在 Int J. Man-Machine Studies, 25, 191-228, 1986。可是在大多数文件编排系统中至少有提供产生数据存储文件的设施。

采用 JASPER 代理程序解决该问题允许用户通过丰富得多的元信息集访问信息。

参照图 1，根据本发明实施例的信息访问系统可以建成信息检索结构的公知格式，诸如连接到因特网的用户-服务器类型结构。

更详细地，用户诸如一家跨国公司可以有多个用户装备个人计算机或工作站 405。这些可以通过用户的用户范围内的万维网（WWW）浏览器 400 连接到用户的 WWW 文件服务器 410。Jasper 代理程序 105 实际上是浏览器 400 的扩展，可以实际驻留在 WWW 文件服务器 410

上。

用户的 WWW 文件服务器 410 以公知的方式连接因特网，例如通过用户自己的网络 415 和路由器 420。业务提供者的文件服务器 425 然后可以通过因特网再通过路由器来访问。

一个文本概述工具 120 和两个数据存储器也驻留在该用户的文件服务器 410 上，或可由该服务器访问，一个数据存储器保存用户文档（文档存储器 430）另一个（智能页面存储器 100）主要保存用于文献收集的元信息。

在基于系统的 Jasper 代理程序中，代理程序 105 自己可以建成诸如公知的网景浏览器的扩展。该代理程序 105 实际上与浏览器 400 成为一体并能够从浏览器 400 提取 W3 页面，该浏览器可以由 Netscape 或由 Mosaic 等提供。

如上所述，在用户-服务器结构中，文本概述器 120 和用户文档都是 Jasper 驻留的用户文件服务器 410 中的文件。可是，Jasper 代理程序 105 可以另外出现在用户的用户范围内。

作为软件代理程序的 Jasper 代理程序通常可以被描述为软件实体，与本地数据一起，或利用支持一个任务或多个任务的本地数据，组成代表用户执行一个任务或多个任务的功能。这些在 Jasper 系统中相关联的任务将在下面进行描述，一个或多个任务可以由 Jasper 代理程序执行。本地数据通常包括来自智能页面存储器 100 和文档存储器 430 的数据，并且由 Jasper 代理程序提供的功能通常包括应用文本概述工具并存储结果、访问和读取、和更新至少一个用户文档的装置，比较关键字集与其它关键字集、或元信息的装置，和对用户发出告警消息的装置。

在优选实施例中，为选择要比较的关键字集的用途 Jasper 代理程序也装备了监视用户输入的装置。

在另一个优选实施例中，Jasper 代理程序装备了涉及第一和第

二关键字集的算法以产生它们之间相似性量度。根据相似性的量度，第一或第二关键字集然后可能由 Jasper 代理程序有效地更新，或可以修改第一或第二关键字集与第三关键字集或元信息比较的结果。

本发明的优选实施例可以根据不同的软件系统构建。例如可以方便地应用面向对象技术。可是，下面描述的实施例中，该服务器将是基于 Unix 并能够运行 ConText™（Oracle 公司提供的一种已知的处理系统的自然语言），和一种 W3 浏览器。该系统通常可以在“C”中实现，尽管该用户可能是潜在的任何能够支持 W3 浏览器的设备。

在下列部分中，讨论 Jasper 代理程序在管理信息中为用户提供的设施。这些可以分成两类，存储和检索。

存储

图 2 和 5 表示当 Jasper 代理程序 105 在智能页面存储器（IPS）100 中存储信息所采用的动作。用户 110 首先找到一个足以感性趣的 W3 页面通过与该用户相关的 IPS100 中的 Jasper 代理程序存储(步骤 501)。该用户 110 然后通过用户选择的 W3 客户 115（现在可用于所有平台上的 Mosaic 和 Netscape 版本）上的菜单选项发送“存储”请求到驻留在用户 WWW 文件服务器 410 中的 Jasper 代理程序 105(步骤 502)。Jasper 代理程序 105 然后邀请用户 110 提供相关注解，该注解也被存储（步骤 503)。通常，这可能是用户对页面感性趣的原因，和可能对于其它用户在决定从 IPS100 中检索哪个页面访问非常有用。(在下面进一步讨论信息共享。)

Jasper 代理程序 105 接着从要求的页面中提取源文本，再次通过 W3 上的 W3 客户 115（步骤 504)。源文本以“超文本”格式提供和 Jasper 代理程序 105 首先去掉超文本链接标记语言(HTML)标签(步骤 505)。Jasper 代理程序 105 然后将文本送到文本概述器诸如“ConText”120（步骤 506)。

ConText120 首先分析文献以确定每句的句法结构（步骤 507)。

ConText 分析器坚固并能够应付出现在英语句子中广泛的句法现象。在对句子层分析之后，ConText120 进入“概念处理”阶段（步骤 508）。提供的实施是：

•信息提取：计算文件内容的主索引，在文本中的概念、事实和定义上索引。

•内容减少：概述的几个层次都可应用，范围从文献主要主题列表到整个文献的大纲。

•论述跟踪：通过跟踪文献的论述，ConText 能够提取特别有关某个概念的文献的所有部分。

ConText120 由用户-服务器结构中的 Jasper 代理程序 105 使用：在分析该文献后，服务器产生应用独立标记版本（步骤 509）。来自使用应用程序接口（API）的 Jasper 代理程序 105 的呼叫则能够翻译该标记。使用这些 API 呼叫，从源文本获得元信息（步骤 510）。Jasper 代理程序 105 首先提取页面文本的概述。概述的大小可以由传递给 ConText120 的参数控制，并且 Jasper 代理程序 105 保证获得 100-150 字的概述。使用另一个对 ConText120 的呼叫，Jasper 代理程序 105 然后驱动来自源文本的关键字集。紧接着，有选择地呈现给该用户通过 HTML 形式 125 加入另外关键字的机会(步骤 511)。以此方式，能够提供对该用户特别相关的关键字，同时 Jasper 代理程序 105 提供可能对更广泛用户群更相关的关键字集。

在该处理结束时，Jasper 代理程序 105 产生关于感性趣的 W3 页面的下列元信息：

•ConText 提供的普通关键字；

•用户的特定关键字；

•该用户的注解；

•页面内容概述；

•文献标题；

•通用资源位置（URL）和

•存储器的数据和时间。

另外参照图 3，Jasper 代理程序 105 然后将这些用于该页面的元信息加入到 IPS100 的文件 130 中（步骤 512）。在 IPS100 中，然后使用该关键字（两种类型）为其它页面建立包含元信息文件的索引。

检索

从使用 Jasper 代理程序 105 的 IPS100 中能够以三种方式检索信息。一种是标准关键字检索设施，而其它两种涉及在代理程序组与它们用户之间共享信息。在下面的部分中将描述每个方式。

当 Jasper 代理程序 105 安装在用户设备上时，该用户提供个人文档：描述通过 W3 获得的用户感性趣信息的关键字集。该文档由 Jasper 代理程序 105 保持，或至少维护以便确定用户对哪个页面潜在地感性趣。

关键字检索

如图 4、6、7 和 8 所示，为直接检索关键字，该用户通过由 Jasper 代理程序 105 提供的 HTML 形式 300 向 Jasper 代理程序 105 提供关键字集（步骤 601）。使用简单的关键字匹配和计分算法，Jasper 代理程序 105 然后检索出保持在 IPS100 中十个最接近匹配的页面（步骤 602）。当页面被存储时，（与那些由 ConText 自动提取的关键字相反）由该用户提供的关键字在匹配处理中可以给予额外加权。该用户可以事先指定检索阈值，低于该值的页面将不显示。代理程序 105 然后利用链接所检索页面和它们的概述的分类列表动态地构成 HTML 格式 305（步骤 603）。与每个所检索页面计分一起也显示由最初用户所做的注解。该页面然后在它们 W3 用户机上呈现给客户（步骤 604）。

“什么是新的”设施

任何用户可以问 Jasper 代理程序“什么是新的？”（步骤 701）。该代理程序 105 然后询问 IPS100 和检索最新存储的页面（步骤 702）。它然后确定这些页面的哪个最符合用户的文档，再次根据简单匹配和计分算法（步骤 703）。然后 HTML 页面被呈现给用户表示链接新存储的最符合用户文档的页面和其它新存储在 IPS 中页面的分类列表（步骤 704），带有所提供的注解。这样提供给该用户浏览新存储的和用户最可能感性趣的页面，和新存储页面的更全面选择（步骤 705）。

用户能够更新 Jasper 代理程序 105 在任何时间通过 HTML 形式保留的文档，该形式允许他从文档中增加和/或删除关键字。以此方式，该用户可以有效选择工作的不同“环境”。由关键字集定义一个环境（context）（那些组成文档的关键字，或那些在检索查询中指定的关键字）和可以把一个环境看成是在给定时间用户感性趣的信息的类型。

由 Jones 在上面的参考文章中提出了对计算机文档编排系统环境中的信息文档编排使用人工记忆模式的想法。如同他在常规文档编排系统环境中指出的，文件系统中的目录与 Jasper 代理程序 105 所检索的页面集之间类似。页面集可以被认为是动态构成的目录，由它被检索出的环境定义。这是在两种意义上的非常灵活的“目录”观念：首先，出现在这次检索中的页面根据环境可以出现在其它检索中；和其次，对于目录没有鲜明的界线：页面在或大或小程度上“位于”目录中取决于它们与当前环境的匹配。在本方案中，在页面上区分信息的方法数量只由信息自身的多样性和丰富性限制。

与其它感性趣的代理程序通信

参照图 8，当通过 Jasper 代理程序 105 将页面存储在 IPS100 中时（步骤 801），该代理程序 105 检查在“本地组”中其它代理程序的文档（步骤 802）。本地组（community）可以是任何预定的组。

如果该页面符合用户文档具有高于某个阈值的分数（步骤 803），消息例如电子邮件消息可以由代理程序 105 自动地产生并被送到有关用户（步骤 804），通知他发现了该页面。

电子邮件的报头可以是例如格式：

Jasper KW:（关键字）

这允许用户在阅读消息内容之前识别其来自哪个 Jasper 系统。优选地，提供关键字列表和用户可以评价该消息涉及信息的相对重要性。消息报头中的关键字随各个用户不同，取决于来自与用户文档中关键字匹配页面的关键字，这样对每个用户兴趣的消息个人化。消息内容本身可以给出进一步信息例如页面标题和 URL，该 URL 存储该页面和存储器提供的页面上的任何注解。

上述 Jasper 代理程序 105 和系统提供在例如 W3 的分布设备中访问有关信息的非常有用方法的基础。在系统中可以作出改变和扩展而不脱离本发明的范围。例如，在相对简单水平上，可以应用改进的检索技术。例如，向量空间和概率模型可以使用，如同 G Salton 在文章“自动文本处理”所描述的，Addison-Wesley in Reading, Massachusetts, USA 于 1989 年出版。

另外，通过提供对元信息的索引而非对关键字索引可以使索引更通用。例如，额外元信息可以是页面存储的日期和该页面的原来位置（Jasper 能够从 URL 中提取的页面）。这些额外索引允许用户（通过 HTML 形式）制定命令类型：

显示所有我于 1994 年从剑桥大学存储的关于人工智能和信息检索的页面。

在另一个替换版本中，可以由 Jasper 代理程序 105 使用辞典来利用关键字的同义词。这减少了当存储页面时按照使用的相同关键字准确输入的重要性。实际上，有可能利用在几个其它区中的辞典，包括代理程序 105 为其用户保持的个人文档。

自适应代理程序

由 Jasper 代理程序 105 使用用户文档以确定有关它们用户的信息，可是功能可以改善。当用户想改变环境（可能从一个任务重新注意另一个任务，或从工作到空闲），该用户文档必须由增加和/或删除关键字重新指定。对于代理程序更好的方案是当用户兴趣随时间改变时改变用户文档。环境的改变可以以两种方式出现：可以从例如工作到空闲短期切换环境。代理程序能够从为用户保存的当前环境列表中识别并改变到新的环境。例如当由用户访问不同信息的新页面时，可以引发此改变。也可以根据用户兴趣发展在环境中由代理程序保持更长期改变。这些改变可以从代理程序对用户的观察中推导出。例如，在自适应代理程序可以应用的公知技术包括遗传算法、反馈学习和根据记忆推理。这种技术在 MIT 的内部报告中公开，1993 年由 Sheth B.和 Maes.p.所著“用于个人化信息筛选的进化代理程序”。

远程和本地信息的集成

Jasper 系统的另一个可能改型是将用户自己的计算机文档系统与 IPS100 集成，以便在 W3 上和在本地设备上发现的信息在顶层对用户呈现同质。然后文件可以由 Jasper 代理程序 105 访问 W3 页面的类似方式访问，解除了面向名称的文档编排系统对用户的限制并对所有类型的本地和远程信息提供内容可寻址的接口。

在 Jasper 系统中群集

JasperIPS100 和所涉及的文献基本上可以称为类集：它是由关键字索引的一组文献。它不同与文献一般与索引远离安放的“传统”类集：该索引（IPS100）实际上指向指定因特网中文献位置的 URL。另外，当存储时，各种附加元信息段挂接在 Jasper 系统中的文献上，例如存储该页面的用户，可以提供给用户的任何注解等等。

Jasper 系统不同于大多数文献类集的重要方面在于已经由用户

将每个文献输入到 IPS100 中，该用户有意识地决定把文献标记为信息段，他或他的同辈在将来发现该信息段有用。这与所保持的元信息一起使 JasperIPS100 成为非常丰富的信息源。

也已经检查了是否公知的信息检索（IR）技术可以有利地应用于 JasperIPS100。特别是，群集的使用已经在研究中。

群集文献

使用公知的 IR 技术，Jasper 的术语-文献阵列可以使用以计算用于在 JasperIPS100 中识别文献的相似性阵列。相似性阵列给出了在存储中所识别文献相似性的量度。对于每对文献计算小块系数（Dice coefficient）。对于两个文献 Di 和 Dj。

2*[Di∩Dj]/[Di]+[Dj]

这里[X]是在 X 中的术语数量，X∩Y 是共同出现在 X 和 Y 中的术语数量。该系数生成 0 和 1 之间的数字。零系数意味着两个文献没有共同的术语，而系数 1 意味着出现在每个文献中的术语组相同。相似性阵列称为 Sim 代表存储中的每对文献的相似性，以便对于每对文献 i 和 j

Sim (i, j) =2*[Di∩Dj]/[Di]+[Dj]

该阵列可以用于自动产生有关文献的群集，使用 Griffiths 等人在“分级凝聚群集方法用于自动文献分类”中描述的分级凝聚群集处理，Journal of Docimentation，40：3，1984 年九月号，175-205 页。在这样处理中，每个文献初始由自己放在群集中和然后俩个这样类似的群集合并成一个更大的群集，然后必须计算大群集与每个其它群集器的相似性。合并过程将持续直到只有单一一个文献群集保留在最高层。

计算群集（与单个文献相反）之间相似性的方法可以改变。对于 Jasper 存储，可以应用“完全链接群集”。在完全链接群集中，来自两个群集的一对文献之间的最小相似性被用作群集相似性。

Jasper 存储得到的群集结构然后可以用于产生三维（3D）在使用 VRML（虚拟现实模型语言）的 Jasper 系统上的前端。（VRML 是公知语言，用于 3D 图形空间或通过全球因特网和万维网内超链接连网的虚拟世界）。

群集关键字

涉及特定 Jasper 文献集出现的关键字（术语）也可以用确切反映上述文献群集技术群集的方法群集：可以构成用于 Jasper 存储中关键字的相似性阵列，该阵列给出在存储中关键字相似性的量度。对于每对文献，计算小块系数。对于两个关键字 Ki 和 Kj，小块系数由下式给出：

$2*[Ki \cap Kj]/[Ki]+[Kj]$

这里[X]是其中出现 X 的文献数量，和 $X \cap Y$ 是其中共同出现 X 和 Y 的文献数量。

可是，一旦计算出用于 Jasper 存储的相似性阵列，当群集该文献时不必群集该关键字。相反有可能用下面描述两个方法利用该阵列。

第一种方法是“文档增强”。在此，可以通过使用那些与用户文档中关键字最相似的关键字增强该用户文档。这样例如，如果词“虚拟”、“现实”和“因特网”是用户文档的一部分但是“VRML”不是，增强的文档可以将 VRML 加入到原有文档（假设 VRML 群集得接近虚拟、现实和因特网）。以此方式，包含 VRML 但不包含虚拟、现实和因特网的文献可以被检索到，尽管它们与非增强文档无关。

图 9 表示关键字 900 的实例网络，该关键字已经根据从当前 Jasper 存储中提取的关键字相似性阵列建立。该算法是直接的：给出一个初始关键字，根据相似性阵列找到四个最相似的词。将这四个词与原有词链接并对四个新词的每个重复该过程。这可以重复许多次（图 9 中三次）。在两个词之间的双线 901 表示两个词都出现在

其它四个最相似关键字中。当然可以对每个涉及词之间相似性程度微粒形信息的链接附加特定的相似性系数。

第二种方法是前摄（proactive）搜索。组成用户文档的关键字可以用于通过 Jasper 以前摄方式搜索与它们兴趣相关的新的 WWW 页面，Jasper 然后可以显示用户可能感性趣的新页面列表而不必用户明确地执行搜索。这些前摄搜索可以由 Jasper 系统以某些给定间隔执行，例如每周。在此群集是有用的，因为文档可以反映多于一个的兴趣。例如考虑下列用户文档：因特网，WWW，html，足球，曼彻斯特，联队，语言学，句法分析，语法学。明显地，在上述文档中显示了三个单独的兴趣，每个分开搜索比起只输入整个文档作为对给定用户的查询有可能产生好得多的结果。根据文献集群集关键字能够使用户 Jasper 代理程序前摄搜索产生查询过程自动化。

当由 Jasper 获得搜索结果时，它们能够被总结并以通常方式与用户文档匹配，以给出与本地保存概述一起的新 URL 排序列表。

对 Jasper 系统的改进

本发明的实施例提供了对 JASPER 系统的改进。现在参照图 10 将描述这些实施例，这些实施例识别在 Jasper 代理程序中的单元，这些单元用于在文献中识别相关关键字，可以改善上面 Jasper 系统的性能。

通过识别两个或更多相互相关的关键字可以增强上述群集技术，例如形成词组的关键字。这些相关关键字然后作为单独术语输入到文献术语阵列。

例如，标准群集技术应考虑到表达式“信息技术”在文献术语阵列中形成两个单独条目，即“信息”和“技术”的单独条目。根据本发明实施例的增强技术应确认词“信息”和“技术”相关和将在文献术语阵列中形成一个单一条目。以一个单一条目替换两个条目能够显著地改变用于在文献之间量度相似性的小块系数值。

例如，考虑下面两段：

1) The people in my company only use the latest information technology when transferring copies of files across our local area network.

2) My company has transferred a lot of people into the latest areas of technology. There is a file on the network with a lot of information in it about the transfers. I also hve a local copy of the file .

明显地，每个段的主题不同，每段具有相同关键字即 "people"，"company"，"latest"，"information"，"technology"，"copy"，"transfer"，"file"，"local"，"area" 和 "network"。

如果关键字 "information" 和 "technology" 和 "local"，"area" 和 "network" 作为独立关键字对待（如同每个标准群集）则对于两个段的小块系数值为 1。如同下面的每个实例，计算使用标准群集技术的该文献术语阵列如下：

	Paragraph1	Paragraph2
People	1	1
Company	1	1
Latest	1	1
Information	1	1
Technology	1	1
Copy	1	1
Transfer	1	1
File	1	1
Local	1	1
Area	1	1
Network	1	1

该阵列表示对两个段有 11 个术语相同和每个段包含 11 个术语。

计算小块系数：

小块系数= (2x11) / (11+11) =1

可是，如果该关键字“information”和“technology”相关在文献术语阵列中形成单一条目，和如果该关键字“local”、“area”和“network”在文献术语阵列中相关形成单一条目，则对于两个段的小块系数重新计算为 0.6。计算如下：

增强的文献术语阵列

	Paragraph1	Paragraph2
people	1	1
company	1	1
latest	1	1
information	0	1
technology	0	1
copy	1	1
transfer	1	1
file	1	1
local	0	1
area	0	1
network	0	1
information technology	1	0
local area network	1	0

该阵列表示对于两个段有六个术语。段 1 具有 11 个术语和段 2 具有 8 个术语，因此：

小块系数= (2x6) / (11) + (8) =12/20=0.6

小块系数 0.6 可以被 认为更精确的反映了两个段主题之间的相似性和区别。

各种段结构和语法结构具有识别关键字集的高概率，这些关键字集以它们的内容作为相似性阵列中的单一条目有可能增强其结果的方式相关。由两个名词组成，或有一个动词跟随一个名词组成的相邻关键字是出现在短语中的普通语法结构类型的实例，因此有可能改善相似性阵列的质量。由一个形容词跟随一个动词是不可能出现在短语的组合，因此被认为不可能增强相似性阵列的质量。

本发明的实施例将包括这种短语结构和语法结构的列表。将检查正在分析的文献文本查找形成这种结构的关键字集的出现。这对识别这些关键字的最初处理是额外的。

在某些情况下有例外，由此被识别为符合特定语法结构的特定关键字集不增强相似性阵列结果。在某些情况下，其它不遵从这些所识别的语法结构之一的关键字集将增强相似性阵列。

因此，本发明的实施例需要在只识别那些具有增强相似性阵列高概率的语法结构与也识别许多具有增强相似性阵列低概率的语法结构之间找到一个折中。

图 10 是在文献中识别相关关键字的 Jasper 代理程序 105 中单元的表示。

输入文本 1000 从 W3 客户设备 115 下载到 Jasper 代理程序 105 中，在那里由第一分析器 1005 分析，“分析器 1”。分析器 1 1005 分析输入文本 1000 的缩写和首字母缩写。

该分析通过比较输入文本 1000 的每个词与缩写和首字母缩写数据库 1010 来实现。分析器 1 1005 将识别出的缩写和首字母缩写加标签。

一旦输入文本 1000 中的缩写和首字母缩写被加上标签，输入文本 1000 然后由分析器 1 1005 再次分析以便将其用空行分割成词群

1010，例如句子，段落，报头（例如 HTML 报头）或项目。

识别缩写和首字母缩写的标签允许分析器 1 1005 的第二次分析处理以区别出现在缩写或首字母缩写结尾的句号和句子结尾的句号。这有助于防止句子中词群 1010 的虚假分离，这可能由在缩写或首字母缩写结尾出现句号引起。

由分析器 1 1005 第二次分析后，该词群 1010 被输入到第二分析器 1020，“分析器 2”。分析器 2 1020 在每个词群 1010 上执行四次操作。

首先，分析器 2 分析词群 1010 中带有不寻常大写的词。这种词通常用于实体名称，例如公司通信网络或计算机系统。例如，可以想象一个公司选择呼叫它的计算机系统之一“Over”。可能在句子中间出现“Over”，在此情况下该词被加标签作为带有不寻常大写的词。该类型的其它变体可能预计包括 OvEr，OveR。被识别为具有不寻常大写的词被标记为“终止列表”忽略(override)。

终止列表包括通常不反映文献信息内容的词的列表。例如，诸如“as”，“is”，“are”，“the”，“they”，“where”，“by”，“my”等。

终止列表可能也包括前缀和后缀的列表。在该例子中运行终止列表以减少带有前缀或后缀的词，变成没有前缀或后缀的基本形式。这被称为词干化，例如“manufacturing”被减少为“manufacture”，“predetermination”减少为“determine”和“preselect”减少“select”。

第二，词群 1010 被与“终止列表”数据库 1025 比较。

第三，不在终止列表中的词和标记为终止列表忽略的词被标记为与文献信息内容有关。

第四，被标记为与文献信息内容有关的每个相邻词对被进一步标记为关键字集，该关键字集可能增强相似性阵列结果。最好，被标记为与文献信息内容有关并由终止列表上的词分开的每对词不被

认为形成相关关键字。

最后，使用将词分类为动词/副词/名词/形容词的字典 1030，根据它们的语法结构识别这些关键字集。这些结构由关键字集中词类型的组合定义，例如，第一个结构可能是由一个动词跟随一个名词而另一个结构可能是由一个名词跟随一个形容词。

在语法结构优选列表范围内的关键字集然后被加上标签，在相似性阵列内容中作为单一条目而不是作为单独的各个条目。

下列列表是被认为可能增强相似性阵列的优选语法结构的列表

相关关键字列表

Word 1	Word 2
noun	noun
verb	noun
adjective	noun
?	noun
?	verb
noun	?
verb	?
adjective	?
?	?

这里“？”代表由 JASPER 代理程序使用不在字典中的词。另外，“？”也可以代表首字母缩写或出现在文献中带有不寻常大写的词。这种词的例子包括 IT，LAN，WAN，xDSL 和 OveR。

在这些例子中，IT 通常用于指“信息技术”，LAN 是“局域网”，WAN 是“广域网”，xDSL 一般是指称为“数字用户线路”技术的一类技术，和 OveR 可以是公司设施的名称例如通信网络。

这些结构不形成确定的列表。其它三个关键字集的结构，例如由两个名词跟随一个形容词，如同在 Local Area Network 的例子中，

也可以被定义。可能增强相似性阵列的四个或更多关键字集结构也可以识别，例如由两个名词跟随一个形容词，该形容词再跟随一个名词，如同在 Asymmetric Digital Subscriber Line 的例子，尽管不如两个或三个关键字集普遍。

这些语法结构提供了在不使用相关关键字与假设每对或三个相邻关键字是相关关键字之间的折中。对于某些主题，特定分类证实对于例如法律文章比对于技术文章更具有好处。因此，可以根据由 JASPER 代理程序 115 所分析的文章类型调整分类。

由于上面详述的每个这些相关关键字作为单一的复合关键字输入到 Jasper 关键字存储中，它们也可以用于增强用户文档的关键字群集技术中。这可以改善由 JASPER 代理程序 115 执行的前摄搜索的质量。这也可以被搜索引擎或类似设备使用，以识别包含相关关键字的文献，这些关键字已经用于定义搜索的目标信息。

该过程不限于英语文献。类似技术可以用于其它语言。

对上述实施例的综合评价

鉴于此中描述的实施例，熟练的收信人将理解其它文本概述器可以用于替代 ConText。例如，ProSum 是由英国电信上市公司在因特网上位于 http: //www.labs.bt.com 的英国电信实验室工作室上提供的概述工具。

尽管所描述的涉及通过因特网的定位信息，本发明的实施例可以对在其它系统上的定位信息有用，例如在超文本的用户内部系统上的文献。

00-03-02

说 明 书 附 图

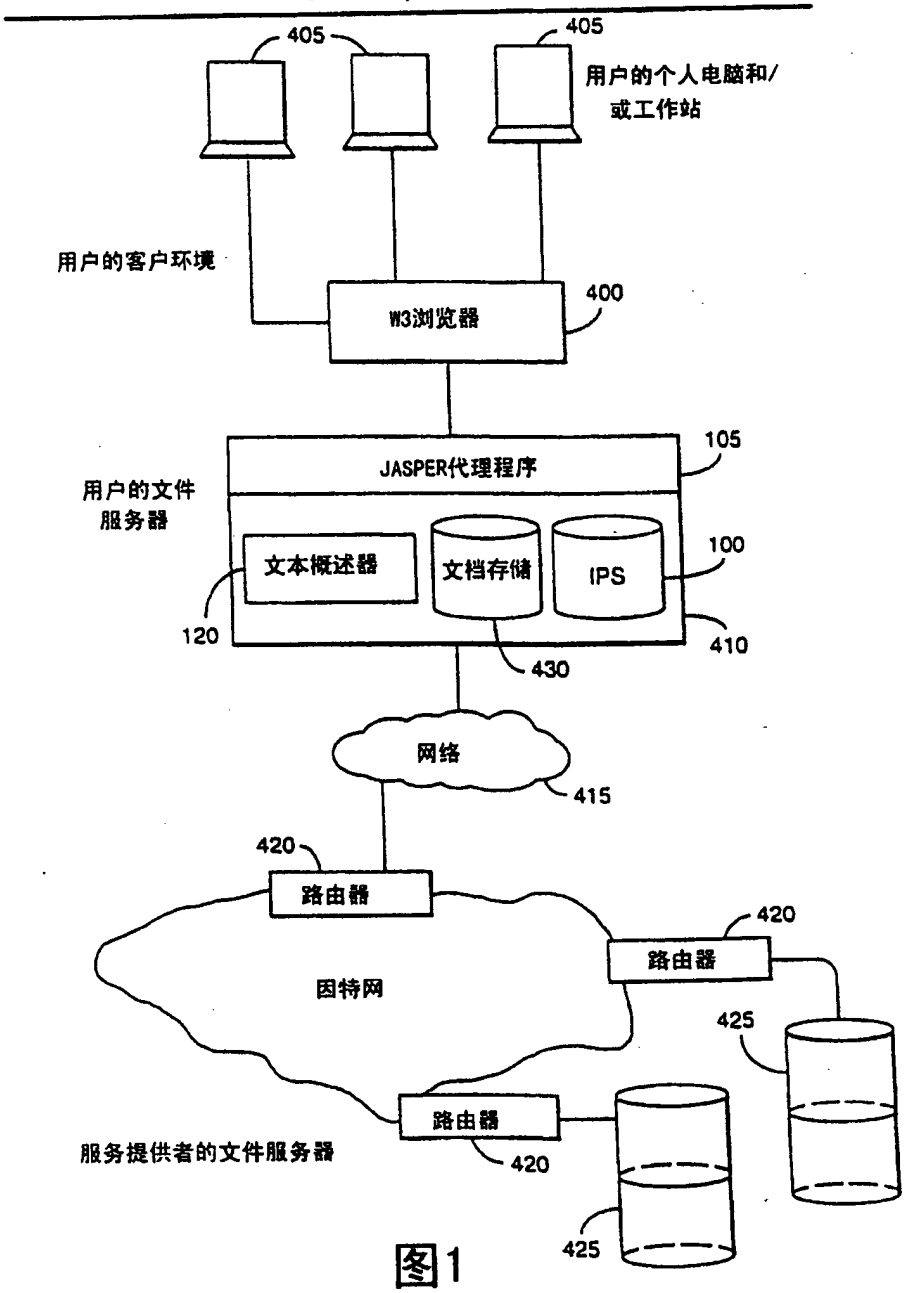


图1

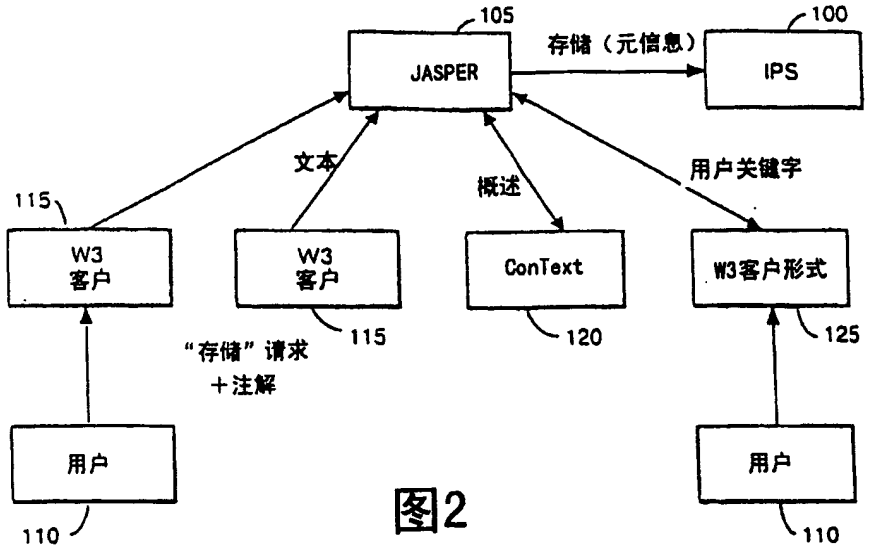

105
JASPER
存储（元信息）
100
IPS
文本
概述
用户关键字
115
W3
客户
W3
客户
ConText
W3客户形式
115
120
125
“存储”请求
+注解
用户
用户
110
110

图2

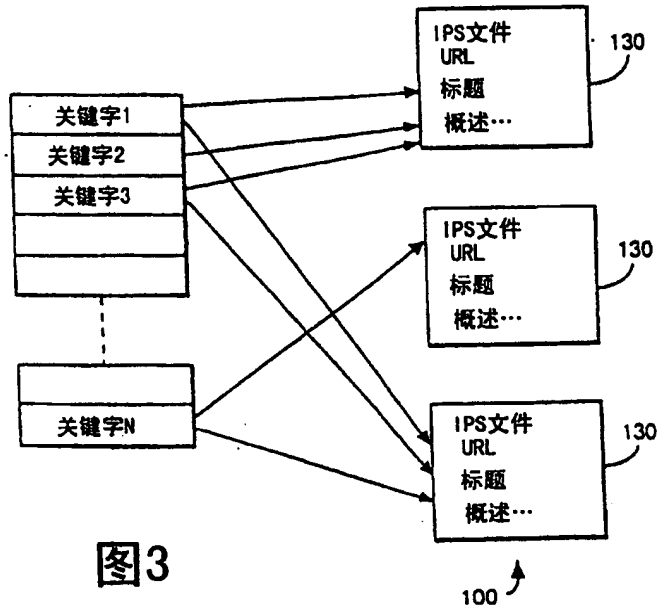

IPS文件
URL
标题
概述…
130
关键字1
关键字2
关键字3
IPS文件
URL
标题
概述…
130
关键字N
IPS文件
URL
标题
概述…
130
100

图3

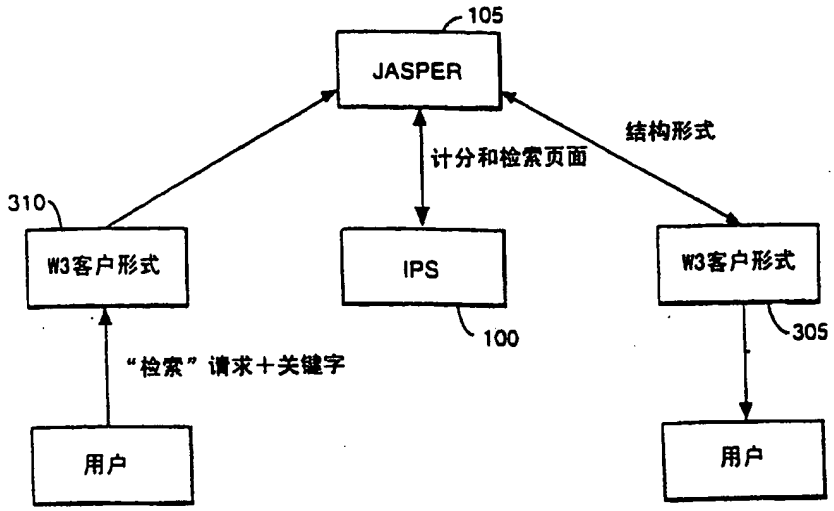

105
JASPER
结构形式
计分和检索页面
310
W3客户形式
IPS
W3客户形式
100
305
"检索"请求+关键字
用户
用户

图4

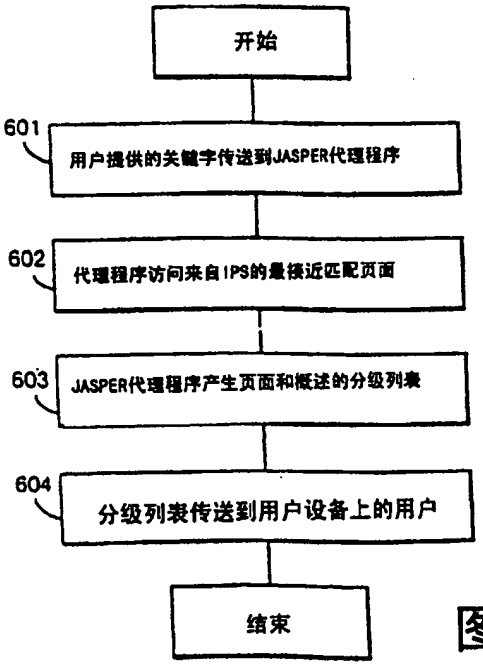

开始
601
用户提供的关键字传送到JASPER代理程序
602
代理程序访问来自IPS的最接近匹配页面
603
JASPER代理程序产生页面和概述的分级列表
604
分级列表传送到用户设备上的用户
结束

图6

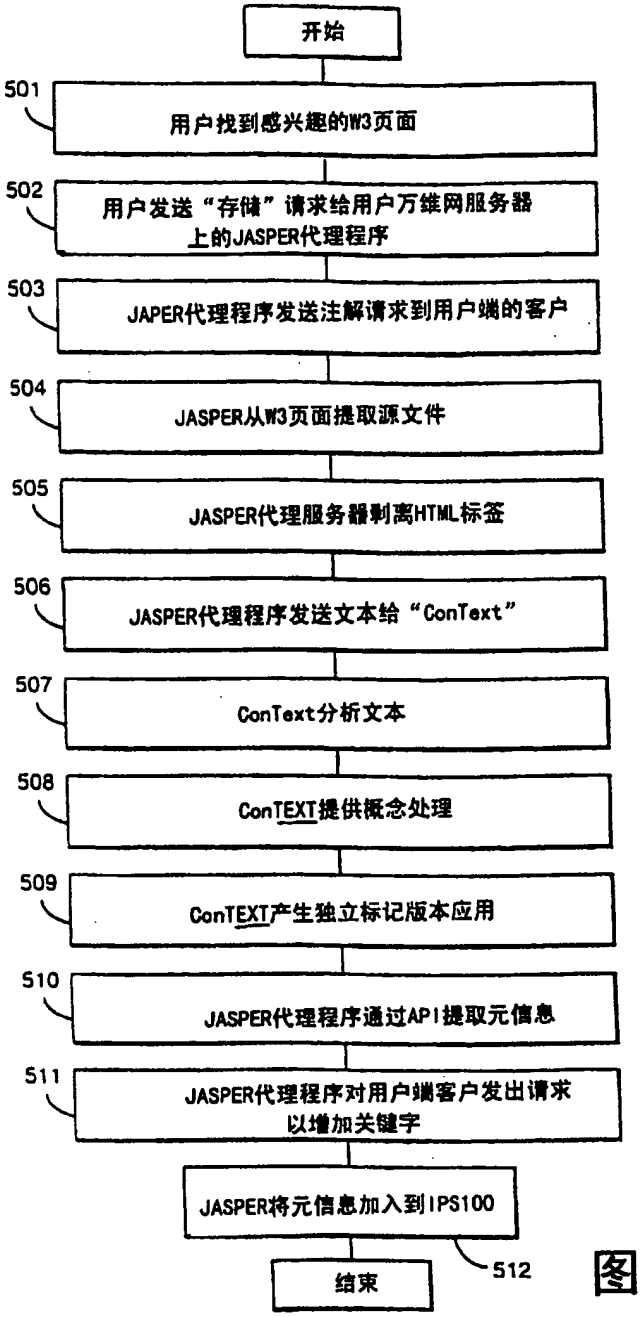

开始
501
用户找到感兴趣的W3页面
502
用户发送“存储”请求给用户万维网服务器上的JASPER代理程序
503
JAPER代理程序发送注解请求到用户端的客户
504
JASPER从W3页面提取源文件
505
JASPER代理服务器剥离HTML标签
506
JASPER代理程序发送文本给“ConText”
507
ConText分析文本
508
ConTEXT提供概念处理
509
ConTEXT产生独立标记版本应用
510
JASPER代理程序通过API提取元信息
511
JASPER代理程序对用户端客户发出请求以增加关键字
JASPER将元信息加入到IPS100
512
结束

图5

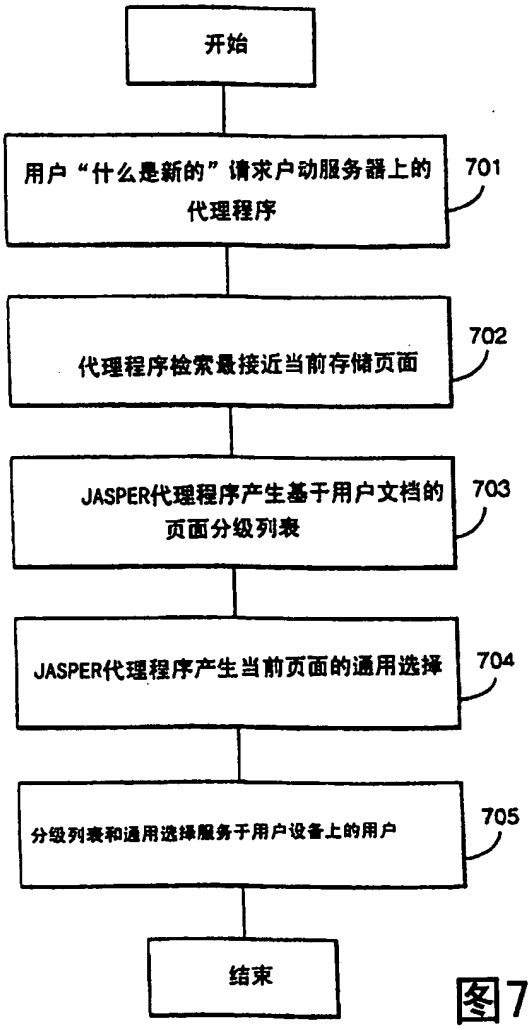

开始
用户“什么是新的”请求户动服务器上的代理程序
701
代理程序检索最接近当前存储页面
702
JASPER代理程序产生基于用户文档的页面分级列表
703
JASPER代理程序产生当前页面的通用选择
704
分级列表和通用选择服务于用户设备上的用户
705
结束

图7

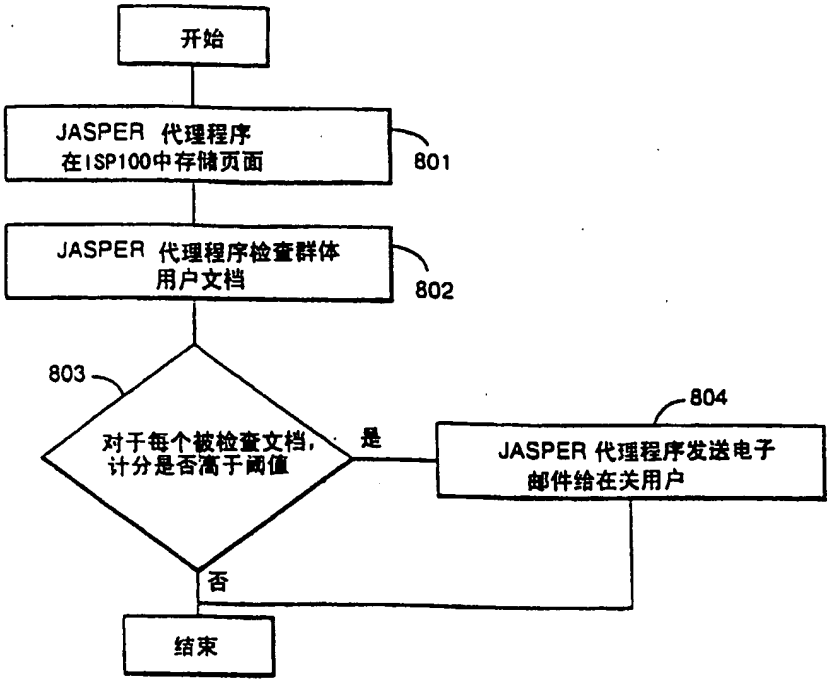

开始
JASPER 代理程序
在ISP100中存储页面
801
JASPER 代理程序检查群体
用户文档
802
803
对于每个被检查文档，
计分是否高于阈值
是
804
JASPER 代理程序发送电子
邮件给在关用户
否
结束

图8

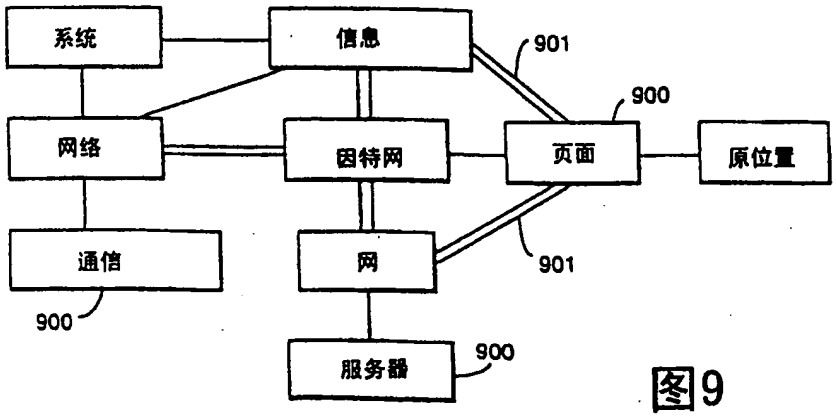

系统
信息
901
900
网络
因特网
页面
原位置
通信
网
901
900
服务器
900

图9

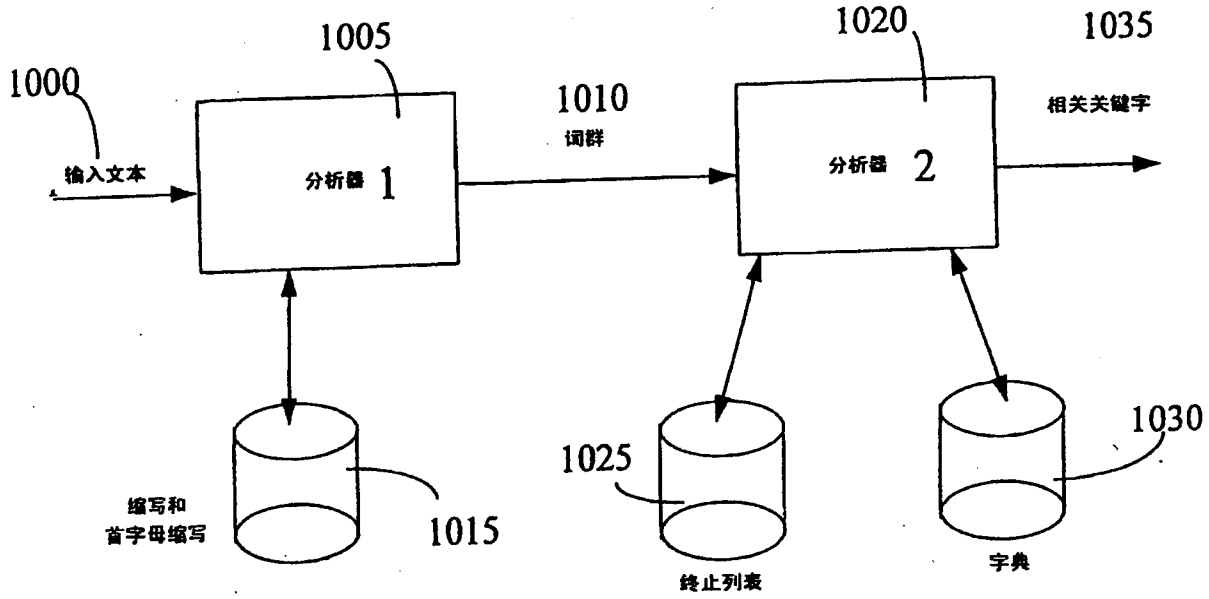

1000
输入文本
1005
分析器 1
1010
词群
1020
分析器 2
1035
相关关键字
缩写和
首字母缩写
1015
1025
终止列表
1030
字典

图10